

Aryan Mishra

AI/ML Engineer | Final-Year CS (Data Science) Undergraduate

Indore, India | +91 82697 13811 | aryanmishraa18@gmail.com

[linkedin.com/in/aryanmdev](https://www.linkedin.com/in/aryanmdev) | github.com/Aryanmishra-dev

Summary

Final-year Computer Science (Data Science) undergraduate and aspiring AI/ML Engineer with hands-on experience designing and shipping three end-to-end AI systems: a RAG evaluation harness, a multi-agent healthcare platform, and a multilingual NLP engine. Core strengths include Python, FastAPI, LangGraph, RAG pipelines, and MLOps tooling with Docker, CI/CD, and MLflow. Demonstrated measurable results including a 100% top-5 retrieval hit rate, 92% agent-routing precision, and 78.1% multilingual sentiment F1, validated with automated test suites and CI quality gates.

Skills

Programming Languages: Python, SQL

AI/ML: PyTorch, scikit-learn, LangChain, LangGraph, Hugging Face Transformers, Natural Language Processing (NLP), Prompt Engineering, Retrieval-Augmented Generation (RAG), Multi-Agent Systems, Model Quantization (ONNX INT8)

Retrieval & Search: ChromaDB, FAISS, BM25, Vector Databases, Hybrid Retrieval, Semantic Search, Cross-Encoder Reranking, Reciprocal Rank Fusion (RRF)

Backend: FastAPI, REST APIs, PostgreSQL, SQLAlchemy, Pydantic, JWT Authentication, Server-Sent Events (SSE)

Cloud & MLOps: Google Cloud Platform (GCP), Docker, Docker Compose, Nginx, GitHub Actions (CI/CD), MLflow, DVC, Git, pytest, Pyright, Bandit

Protocols: Model Context Protocol (MCP)

Projects

RAG EvalForge — Production-Grade RAG Evaluation Harness (*Personal Project*) *April 2026 – August 2026*

- Benchmarked four chunking strategies—fixed, recursive, sentence, and semantic—on a hand-labeled question-answer dataset, achieving a 100% top-5 retrieval hit rate and 0.858 mean reciprocal rank (MRR) with recursive chunking.
- Diagnosed semantic-chunking over-segmentation through threshold sweeps, improving top-5 retrieval accuracy by 22%.
- Integrated dense retrieval with ChromaDB, lexical retrieval with BM25, and cross-encoder reranking using RRF, reducing retrieval latency by 35%.
- Built an LLM-as-a-judge evaluation framework to score retrieval faithfulness, answer correctness, and response relevance across multiple embedding models.
- Implemented continuous integration quality gates with 108 pytest cases, 89% test coverage, strict Pyright checks, and Bandit security scanning on every pull request.
- Developed an MCP server exposing ingestion, benchmarking, and model-registry tools for clients including Claude Desktop.

MedAgent AI — Multi-Agent Healthcare Conversational Platform (*Personal Project*) *November 2025 – March 2026*

- Designed a LangGraph supervisor to route user queries across four specialized agents—Research, Chat, Recommendation, and Summarization—achieving 92% intent-routing precision across 500+ test scenarios.
- Built a domain-specific retrieval-augmented generation (RAG) pipeline over a healthcare knowledge base with persistent memory, improving multi-turn context retention by 30%.
- Developed a FastAPI and PostgreSQL backend with JWT authentication, server-sent events (SSE) token streaming, and asynchronous task processing with Celery and Redis.
- Containerized the platform with Docker Compose and load-tested the Nginx deployment with 500+ concurrent sessions.

Multilingual ABSA — Cross-Lingual Aspect-Based Sentiment Engine (*Personal Project*) *June 2025 – October 2025*

- Built a multilingual aspect-based sentiment analysis (ABSA) engine for English, Hindi, and Hinglish reviews, achieving 78.1% F1 in English and 67.8% F1 in Hindi.
- Optimized inference with ONNX INT8 quantization, achieving 185 ms p95 latency while preserving production-ready model performance.
- Implemented an inference fallback chain from INT8 to FP32 to rule-based processing, achieving 100% request availability during testing.
- Tracked the end-to-end machine learning pipeline—including preprocessing, training, quantization, and model export—with DVC and MLflow, and added drift monitoring.
- Shipped a Streamlit web application supporting single-record and batch-CSV predictions with live monitoring dashboards.

Education

Indore Institute of Science and Technology

Expected May 2027

Bachelor of Technology in Computer Science Engineering (Data Science)

CGPA: 7.41/10 (74.1%)

Certifications & Achievement

- NVIDIA Deep Learning Institute — Anomaly Detection for Predictive Maintenance and Computer Vision (2025)
- Google Cloud — Generative AI Fundamentals (2026); Large Language Model Implementations
- Anthropic — Model Context Protocol, Claude 101, Claude Code 101, Introduction to Agent Skills
- Cisco — Apply AI: Analyze Customer Reviews; Introduction to Modern AI (2026)
- Kaggle Bronze Medalist — March Machine Learning Mania 2024